

Mutable Transcripts

Mitigating Context Pollution through Editable
Conversation State

Dan Barry & Andrew Hines

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

What was U.S. GDP in 2021?

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

What was U.S. GDP in 2021?

About \$23.32 trillion.

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

What was U.S. GDP in 2021?

About \$23.32 trillion.

What was the largest category?

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

What was U.S. GDP in 2021?

About \$23.32 trillion.

What was the largest category?

Personal consumption expenditures and ...

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

What was U.S. GDP in 2021?

About \$23.32 trillion.

What was the largest category?

Personal consumption expenditures and ...

Sorry! I meant 2022.

Chat remembers **everything**, including what you no longer mean.

Most LLM chat is an append-only transcript. Every correction becomes another turn in the context supplied to the model.

~~What was U.S. GDP in 2021?~~

~~About \$23.32 trillion.~~

What was the largest category?

~~Personal consumption expenditures and ...~~

Sorry! I meant 2022.

For 2022: U.S. GDP was about \$25.46T; personal consumption expenditures remained the largest category, at about \$17.56T.

CONSEQUENCE

The correction arrives several turns later. Earlier assumptions and dependent answers still remain in context. The chat history is cluttered and requires user effort to parse later

Accumulated corrections turn a useful thread into a noisy one.

01

Outdated facts

A resolved misunderstanding continues to occupy the transcript.

02

Conflicting constraints

New direction competes with what was asked earlier.

03

Irrelevant detours

Abandoned topics can influence later model responses.

Accumulated corrections turn a useful thread into a noisy one.

01

Outdated facts

A resolved misunderstanding continues to occupy the transcript.

02

Conflicting constraints

New direction competes with what was asked earlier.

03

Irrelevant detours

Abandoned topics can influence later model responses.

The model is not simply “forgetting”—it lacks a reliable signal for what should be forgotten.

Existing edits don't revise the conversation's state.

STANDARD CHAT

Append or branch

Edit the latest prompt-or create a divergent branch. Earlier context remains, including obsolete or corrected information.

— — —



MUTABLE TRANSCRIPTS

Revise the state

Ask in natural language to update the history itself, then continue from the revised record.

— — —

Make the transcript an editable representation of conversational state.

$$T' = f(T, e)$$

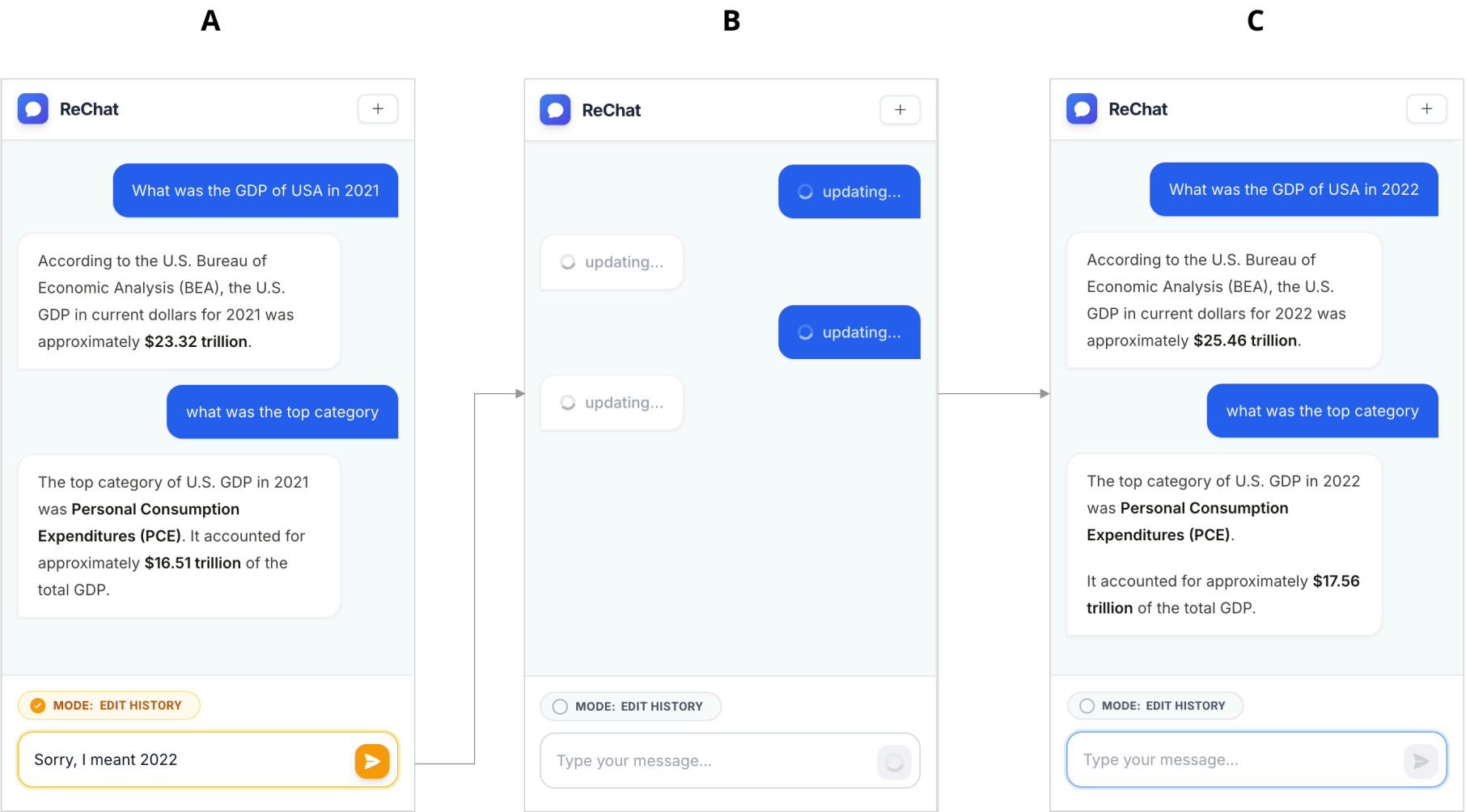
T' revised transcript

T original transcript

e natural-language edit request

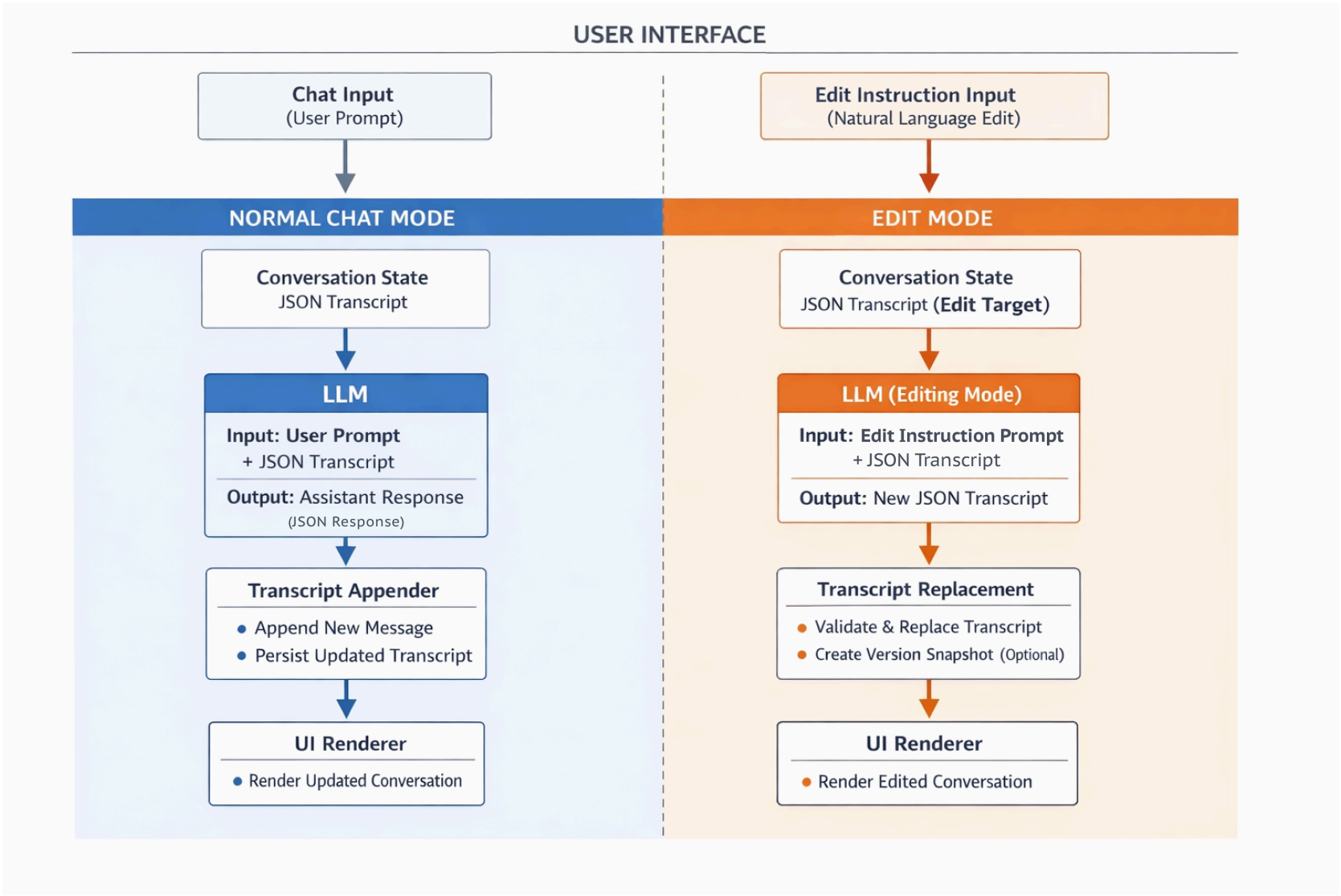
The user describes the change; the system rewrites affected turns so the retained context reflects current intent.

One lightweight mode change. A different relationship to history.



Enter **Edit History** → issue an edit instruction → inspect the revised conversation.

Two modes, one familiar chat interface.



Default mode appends a new turn. **Edit History mode** transforms the full transcript and replaces the displayed state.

History becomes a surface for intent management.



Retroactive correction

Update an earlier mistake and propagate its implications downstream.

STUDY TASK

Correct the country and year in a GDP research conversation. *(year change)*



Context pruning

Remove an irrelevant detour without restarting the whole conversation.

STUDY TASK

Remove an outdated topic that is steering the discussion. *(vegetarian option)*



Constraint injection

Apply a new requirement consistently across the entire discussion.

STUDY TASK

Add new requirements that must apply across the conversation. *(paris trip)*

Does direct revision improve the experience of correcting chat?

17

participants

2

within-subject conditions

3

revision tasks

Standard chat vs. Mutable transcript × correction constraints pruning

For each conversation task, participants rated 5 metrics:

01
Transcript cleanliness

02
State clarity

03
Confidence

04
Restart intent

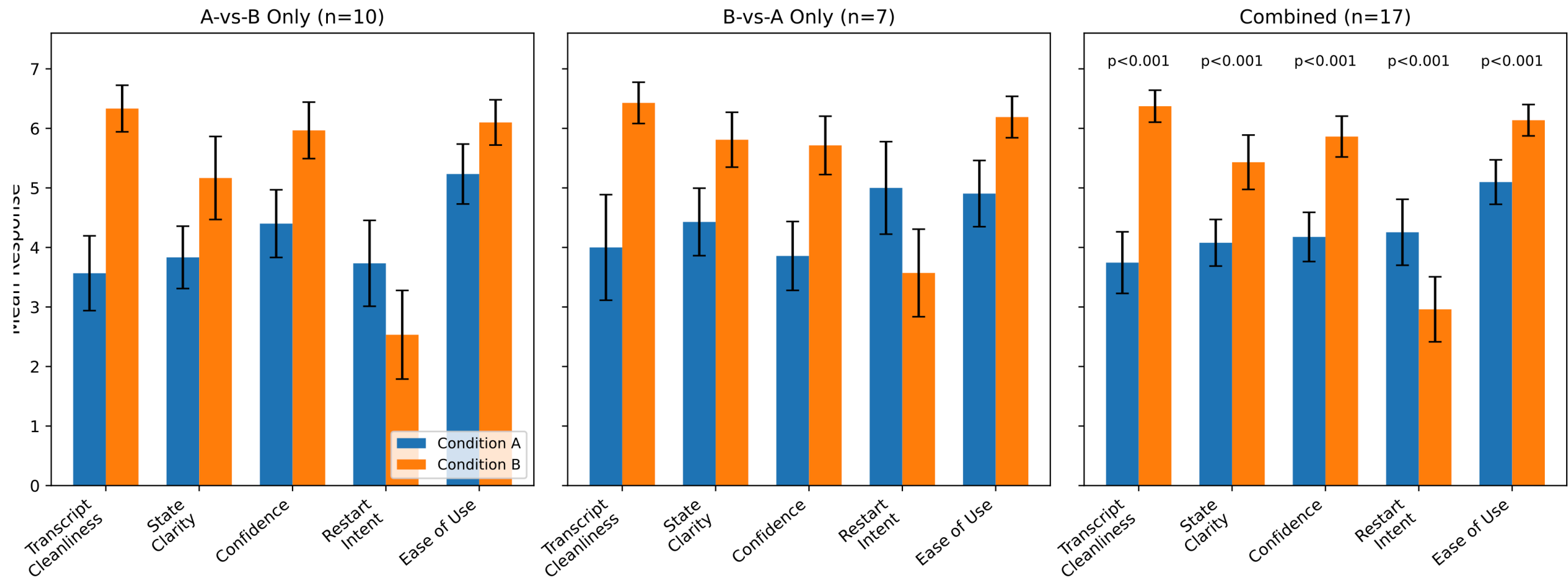
05
Ease of use

For each conversation task, participants rated five measures.

1. **01 Transcript Cleanliness**
"This conversation thread is clean and free of outdated information that could mislead future chatbot responses."
2. **02 State Clarity**
"It is easy to understand what assumptions or constraints are currently in effect within the conversation."
3. **03 Confidence**
"I am confident the chatbot will follow the updated constraints going forward in a long conversation."
4. **04 Restart Intent**
"If this were a real interaction with a chatbot, how likely would you be to start a new conversation instead of continuing with this conversation."
5. **05 Ease of Use**
How easy was it to apply your corrections?

Mutable transcripts outperformed standard chat on all five measures.

Condition Comparison Across A-vs-B, B-vs-A, and Combined



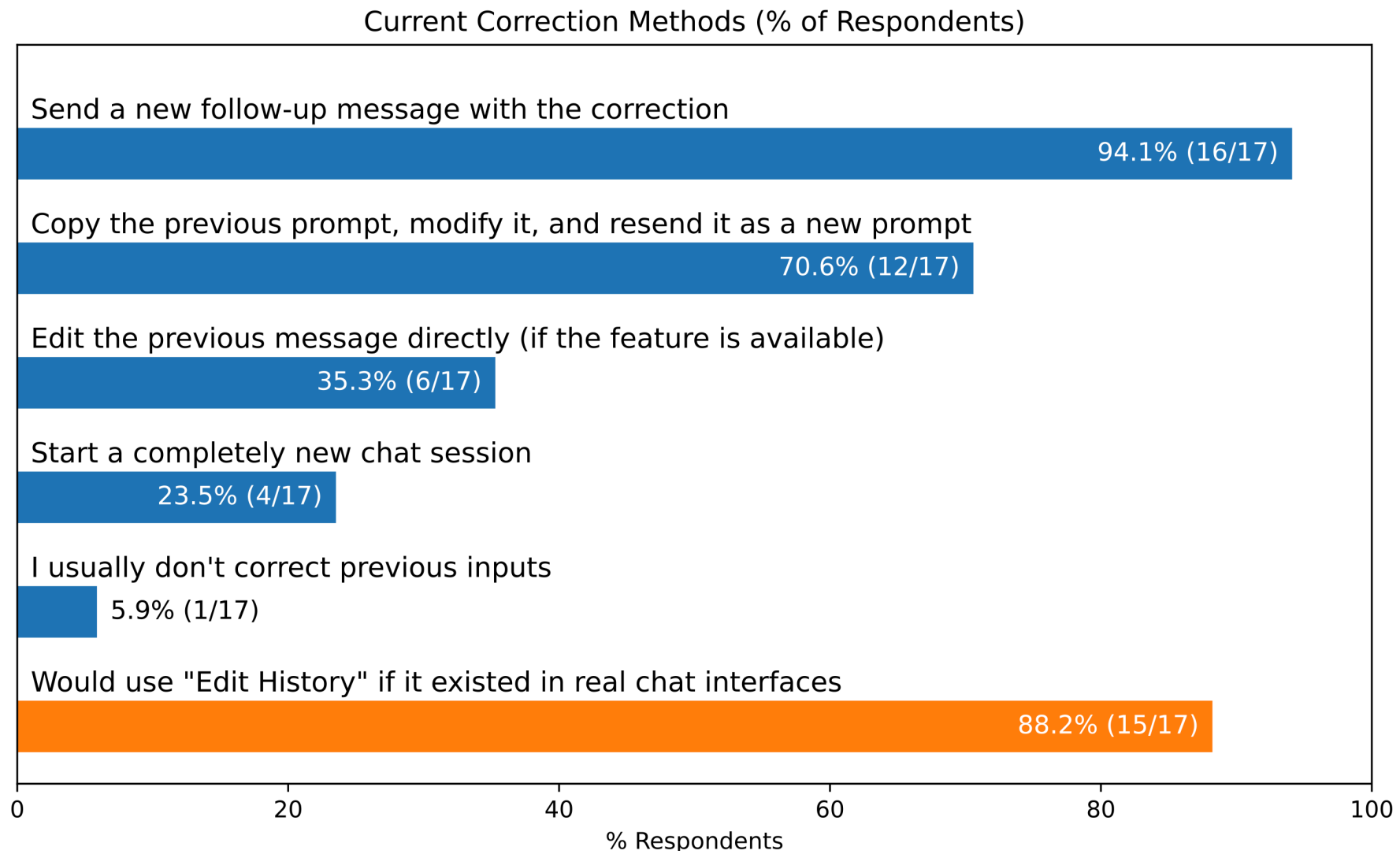
All comparisons $p < 0.001$ Lower restart intent is better.

Revision removes obsolete context instead of piling it on.

Scenario	Method	Turns	Tokens	Obsolete Tokens	Obsolete %
Retroactive Correction	Baseline	8	422	236	56%
	Mutable	4	169	—	—
Constraint Injection	Baseline	6	608	353	58%
	Mutable	4	479	—	—
Context Pruning	Baseline	10	712	655	92%
	Mutable	4	66	—	—
Mean	Baseline	8	581	415	71%
	Mutable	4	238	—	—
Turn/Token Reduction		50%	59%	—	—

A paired transcript analysis of representative correction, constraint, and pruning scenarios.

People already correct chat—just inefficiently.



Editable history is powerful-but it must be legible and accountable.

Potential value

- Cleaner, current conversational state
- Less need to abandon productive threads
- User control without abandoning chat

Design challenges

- Show what changed-and why
- Prevent loss of useful context
- Support previews, undo, and version history
- Respect audit-sensitive settings

When intent changes, history should be able to change too.

Mutable transcripts preserve the conversational interface while allowing users to keep its context aligned with what they mean now.

Source Code: github.com/QxLabIreland/ReChat